Noname manuscript No.
(will be inserted by the editor)

Comparison of Local Explanation Techniques for CNN Models

Ya Liu

Received: date / Accepted: date

Abstract Deep neural network models usually have a high accuracy. However, when it comes to application in practice, duo to complexity of these models, they are often blamed for lack of interpretability. Interpretability is the degree to which a human can understand the cause of a decision. Thus, there is a tension between accuracy and interpretability. In response, numerous explanation techniques are designed to help users understand and interpret these models, but it still remains unclear that when one technique is preferable over another. To approach this problem, this paper analyzes three explanation techniques, DeepLIFT, SHAP (DeepExplainer, GradientExplainer, KernelExplainer) and LIME, and compares their explanations in terms of same images and convolutional neural network models. We can infer that DeepExplainer corresponds with human tuition, and that there is an issue of instability in GradientExplainer and a problem of unstablity in LIME.

Keywords DeepLIFT · SHAP · LIME · Interpretability

1 Introduction

Users need to understand why a model makes a decision and decide whether to trust its prediction. Some models are intrinsically interpretable but not very powerful, e.g., linear models, whereas deep neural network models are powerful but too complex to interpret. Recently, various explanation techniques are proposed to interpret models. These techniques can be classified into two categories, local explanation and global explanation. Local explanation aims to explain each prediction a model makes, while global explanation aims to explain the whole model. Since this paper focuses on local explanation, three

Ya Liu
George Washington University
Tel.: 2022509249
E-mail: yliu2@gwu.edu

techniques are selected, DeepLIFT, SHAP, LIME.

In the following sections, this paper briefly introduces these three techniques, builds two convolutional neural network(CNN) models, and compares explanations of three techniques based on these models and same inputs.

2 Three Explanation Techniques

In this section, I briefly discuss the core ideas behind these three explanation techniques.

2.1 DeepLIFT

DeepLIFT (Deep Learning Important FeaTures) is a gradient based algorithm [1]. This kind of algorithm uses gradient to represent feature importance because gradients can show how the model's prediction changes as the feature changes. However, one limitation of gradient based algorithms is gradient saturation problem, which means the feature has taken a value which has the most impact on the score globally, but the gradient fails to capture this importance [2]. In response, DeepLIFT is proposed to solve this problem. Instead of using gradient, it utilizes slope to represent feature importance. Gradient describes how y changes as x changes at the point x, while slope describes how y changes as x differs from the baseline. In practice, we usually use the mean of the data set or an array of zeros as baseline.

2.2 SHAP

SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model [4]. It argues that six explanation methods are all transformations of additive feature attribution methods. Unified with Shapley value, a concept derived from cooperative game theory and computed to measures each person's contribution among coalition, various explainers are proposed in terms of different kinds of models. For example, TreeExplainer explains the output of ensemble tree models; AdditiveExplainer computes SHAP values for generalized additive models. Among those explainers, three of them (KernelExplainer, DeepExplainer, GradientExplainer) which can be used or specifically designed for deep neural network models are selected.

KernelExplainer has no assumption about models, and is a combination of Linear LIME and Shapley values. GradientExplainer and DeepExplainer are designed for deep neural network models. GradientExplainer is an extension of integrated gradients, whereas DeepExplainer is an enhanced version of the DeepLIFT algorithm.

2.3 LIME

LIME (Local Interpretable Model-Agnostic Explanations) a technique to explain the predictions of any machine learning classifier [3]. The core idea behind is to approximate the underlying model by an interpretable one (e.g., linear model). In terms of images explanation, it first divides an image into interpretable components (contiguous superpixels). Then it generates a data set of perturbed instances by turning some of the interpretable components "off". For each perturbed instance, we get the probability that an object is in the image according to the model. We then learn a simple (linear) model on this data set, which is locally weighted. Finally, we present the superpixels with highest positive weights as an explanation, graying out everything else.

3 Two CNN Models

I build two CNN models to be explained. One is trained on handwritten digits image dataset from MNIST. I also finetune VGG16 on fashion product dataset. This dataset includes 143 classes, and I only choose five of them with each containing about 1000 samples. By simplifying this dataset, I try to develop a model with high accuracy and meanwhile reduce the influence of errors on the performance of explanation techniques. Both models have accuracy over 0.95, and details about these two models can be found in my GitHub repository.

4 Comparison of Three Explanation Techniques

4.1 Experiment Setup

Three explanation techniques are installed using pip in Python. It is worthwhile to notice that the version of Tensorflow and Keras are different (DeepLIFT: Tensorflow == 1.15.2, Keras == 2.3.1; SHAP: Tensorflow==2.3.0, Keras == 2.4.3; LIME: Keras == 2.4.3).

4.2 Results Evaluation

In order to perform a clear and straightforward comparison, I randomly select two digit images, four casual shoes images, and four flip flops images, and compare their explanation using these three techniques.

Figure 1 illustrates explanations of these methods on handwritten digit images with the first column displaying original images. For DeepLIFT, the darker the area is, the more important that area is. For DeepExplainer and GradientExplainer, red areas mean that these areas increase the probability of this image being in this class, while blue areas indicates that these areas decrease the probability of this image being in this class. Similarly, for KernelExplainer,

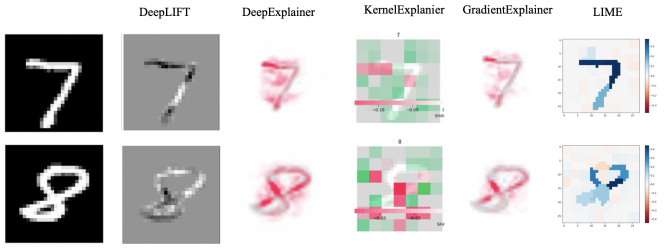

Fig. 1 Comparison Based on Handwritten Digit Images

green areas increase the probability of this image being in this class, and red areas decrease this probability. For LIME, blue areas are important features to classify this image into this class, while red areas are components that make this image less likely in this class.

Interpretation of digit seven and eight can be similar and meanwhile be slightly different. For example, Deeplift and LIME think the northeastern corner of this seven digit is an important part, while DeepExplainer and GradientExplainer do not spot this area. Besides, DeepLIFT puts emphasis on the bottom part of digit eight, while the others leave this part unimportant. On the other hand, all of them recognize that the middle part of digit eight is crucial, and that the beginning and ending parts of digit seven are of great importance. Additionally, DeepLIFT can only select components that have huge impact on prediction whereas other methods can also figure out the positive and negative relationships between these components and predictions. We can also notice that KernelExplainer does not perform well since its segmentation and focus are unreasonable. With these observations, it is still difficult to tell which method is preferable over another because it is subjective to decide whether a part is important and there is no metrics designed to measure goodness of explanation methods.

Figure 2 and figure 3 shows explanations of these methods on casual shoes and flip flops. We can conclude that DeepExplainer can spot shoelaces and heels that are key differences between casual shoes and flip flops. We can notice some problems in other explanation methods. We can see that data points are sparse in explanations of GradientExplainer so there may be a problem of sparseness. The first three explanations in fifth column of figure 3 capture key features, e.g., heels and shoelaces. However, the last explanation loses track of important features. Thus there may be an issue of instability in LIME. In addition, from forth column of figure 2, we can infer that KernelExplainer sometimes looks at background. It seems that DeepLIFT outlines the product in an image instead of filtering out unnecessary parts.

It can be infered from figure 4 and figure 5 that explanation of DeepExplainer is in correspondence with human tuition. When DeepExplainer tries to explain

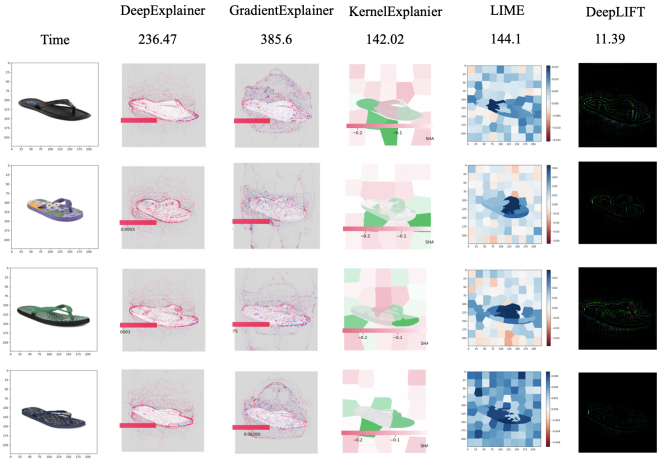

Fig. 2 Comparison Based on Flip Flops

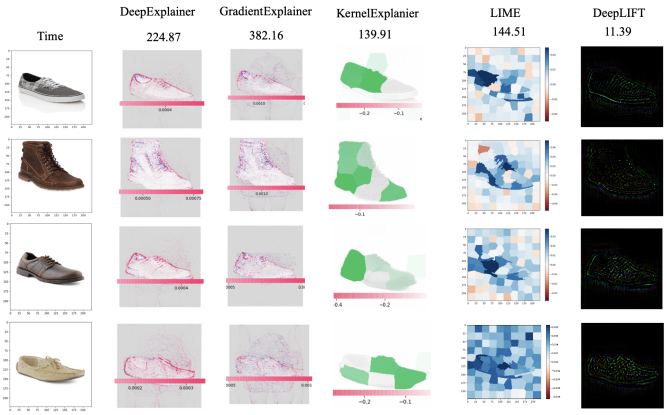

Fig. 3 Comparison Based on Casual Shoes

why digit four image is in digit four class, the red area on top indicates that this area makes this image more "four". But when it tries to explain why this image is not in digit nine class, this area turns to be blue, which means that this area makes this image less "nine". It is significantly reasonable since this is a key part that differentiate digit four from digit nine. We can also see this pattern for digit nine image. The first column of figure 5 shows the original images, and the rest explain why this image is in this class or not with order backpacks, belts, briefs, casual shoes, and flip flops. The first image belong to flip flops class, and the fifth column explains why this image is not casual shoes. We can notice that the blue area is located at the end of these shoes, and this reveals that missing of this part decreases the probability of this image being casual shoes, or we can say, makes it less "casual shoes". On the other hand, the sixth column shows why this image is in flip flops class.

Fig. 4 Explanaion of DeepExplainer on Digit Four and Digit Nine

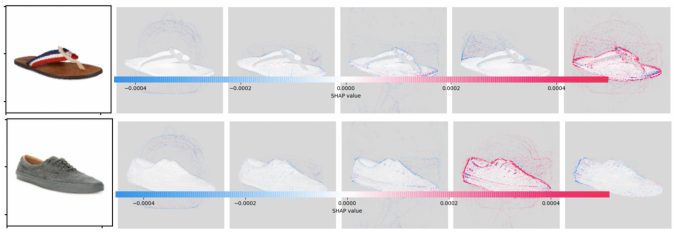

Fig. 5 Explanaion of DeepExplainer on Casual Shoes and Flip Flops

Similarly, we can see why the last image is casual shoes instead of flip flops. It is interesting to notice that the model captures the top line of casual shoes including shoelaces as important features. When it tries to explain why this is not flip flops, it focuses on the left side of shoes. Since shoelaces and heels are key features to differentiate casual shoes and flip flops, DeepExplainer can explain a model to a degree that is in correspondence with human tuition.

Another interesting finding about DeepExplainer is that its explanation reflects performance of a model. We can see from Figure 6 that, after I rotate the original image with different degree, the explanation changes because prediction of model changes. For the second and forth rows, these images are classified into briefs.

5 Conclusion

This project aims to compare the performance of three local explanation methods, DeepLIFT, SHAP, and LIME on two CNN models, thus to conclude which method is preferable over another. First I train two CNN models on handwirtten digit dataset and fashion product dataset. Then I compare them based on same images. We can conclude that there may be a problem of sparseness in GradientExplainer, and an issue of instability in LIME. KernelExplainer tends to recognize background as important parts. We can also infer that explanations of DeepExplainer correspond with human tuition.

Acknowledgements I would like to thank my professor, Amir Jafari, for helpful guidance and feedback. I could not have done this work without his help.

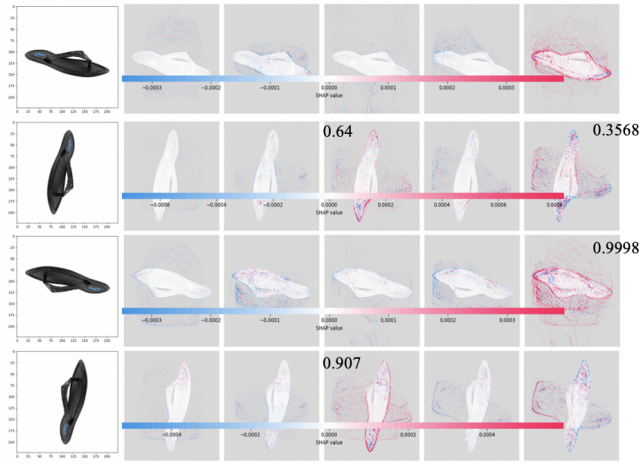

Fig. 6 Explanation of DeepExplainer After Rotating an Image

References

1. Shrikumar, Avanti and Greenside, Peyton and Kundaje, Anshul, Learning important features through propagating activation differences, arXiv preprint arXiv:1704.02685, 2017
2. Mohammadreza Salehi, A Review of Different Interpretation Methods in Deep Learning (Part 2: Pixel-wise Decomposition, DeepLIFT, LIME)
3. Tulio Ribeiro, Marco and Singh, Sameer and Guestrin, Carlos, " Why Should I Trust You?": Explaining the Predictions of Any Classifier, arXiv, arXiv–1602, 2016
4. Lundberg, Scott M and Lee, Su-In, Advances in Neural Information Processing Systems, 4765–4774, Curran Associates, Inc., 2017